Enhancing the Interoperability of iSimp by Using the BioC Format

Yifan Peng[1, *], Catalina O Tudor[1, 2], Manabu Torii[1, 2], Cathy H Wu[1, 2], K Vijay-Shanker[1]

[1]Department of Computer and Information Sciences, University of Delaware, Newark, DE,
[2]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE

*Corresponding author: Tel: 302 831 8496, E-mail: yfpeng@udel.edu

Abstract
This paper reports the use of the BioC format in our sentence simplification system, iSimp, so that it could be seamlessly used in text mining pipelines. iSimp is designed to simplify complex sentences commonly found in the biomedical text, therefore bringing benefits to existing text mining applications that rely on the analysis of sentence structures. By adopting the BioC format, we aim to make iSimp readily integrable in various applications in this domain. To examine the utility of iSimp with BioC, we designed and implemented a rule-based relation extraction system that uses iSimp as a preprocessing module and BioC format for data exchange. Evaluation on the BioNLP-ST 2011 GE task training corpus showed that, with sentence simplification provided by iSimp, the F-value of the phosphorylation extraction increased 3%. The iSimp corpus previously used for the evaluation of simplification and the GE task corpus used in the current study have been converted into the BioC format and made publicly available[1].

Introduction
The syntactic complexity of the biomedical text often poses a major challenge in designing and applying Natural Language Processing (NLP) systems on scientific articles. One possible approach to address this issue and improve the performance of NLP systems (e.g., relation extraction systems) is to simplify the complexity of the sentences themselves prior to using them as input in the NLP systems. For this purpose, we had previously developed iSimp [1], a sentence simplification system. Used a preprocessing module that simplifies the input text, iSimp has a potential to enhance existing text mining applications in the biomedical domain. In order to make iSimp readily integrable in various applications, we have adopted the BioC format, a simple, yet robust, XML format to share text documents and annotations [2].

We report in this paper how BioC is used with iSimp. One of the contributions of this work is a BioC tag set for annotating iSimp outputs. Sentence simplification is a task that there is no

[1] http://research.dbi.udel.edu/isimp/corpus/

standard scheme for annotating simplification results. We define a BioC tag set to share and compare simplification annotation results from various simplifiers.

A second contribution of this work is a mechanism, using the BioC framework, to encode simplified sentences in the corpora. A factor that makes integration of iSimp with BioC format distinct, compared to many NLP tasks, is that the annotation can include sequences and words that are not from the original text. This is because iSimp produces new sentences together with annotation of the simplification constructs in the original text. Thus, the proposed mechanism allows simplified sentences to be included in a BioC annotation file and be treated as part of the original collection for further processing in the NLP pipeline.

A third contribution of this work is the iSimp corpus [1], which consists of 130 Medline abstracts and is annotated with six simplification constructs. We converted the corpus into BioC format and made it public available to be used for evaluation of different simplification systems.
In order to show the wide applicability of iSimp, we examined its impact on event extraction. We designed and developed a simple rule-based relation extraction system. We showed that with sentence simplification provided by iSimp, the performance of the relation extraction system improves. We also present how iSimp can be utilized with BioC, by enabling both iSimp and the relation extraction system use BioC format. This makes the module integration seamlessly. For this, we report another contribution of this work, namely the conversion of the BioNLP-ST 2011 GE corpora into the BioC format and its public availability.

Methods

In this section, we describe iSimp, the relation extraction, the corpus used in our evaluation, and how the BioC format is used to facilitate an easy I/O exchange between these components.

iSimp

iSimp is designed to reduce the sentence syntactic complexity. To illustrate the usefulness of our sentence simplifier, iSimp, consider the following complex sentence:

> E1. A third genetic linkage to disease is alpha-synuclein, a protein that is heavily phosphorylated in Lewy bodies and Lewy neuritis, the pathological hallmarks of PD. (PMID-22342821)

As shown in this example, the major syntactic constructs that we considered for simplification are: coordination (e.g., ‘‘Lewy bodies and Lewy neuritis’’), relative clause (e.g., ‘‘a protein that is heavily phosphorylated in …’’), and apposition (e.g., ‘‘alpha-synuclein, a protein that is …’’ and “Lewy bodies and Lewy neuritis, the pathological hall marks of PD”). For a more detailed description of iSimp, as well as its challenges (attachment ambiguities, boundary detection, and nested constructs), we refer the reader to [1].

iSimp identifies the various types of simplifications (coordinations, relative clauses, appositions, etc.) and breaks the complex sentence into multiple simple sentences. Here we only show two examples of simplifying (E1):

E2. Alpha-synuclein is heavily phosphorylated in Lewy bodies.
E3. Alpha-synuclein is heavily phosphorylated in Lewy neuritis.

We made iSimp available as an online tool[2], and adopted the BioC format as its input/output format. Figure 1 shows the workflow of the system. iSimp first tokenizes the text and then it splits each sentence into a sequence of non-overlapping chunks. The detection of various simplification constructs is based on the chunks, and from these, iSimp then generates simplified sentences.

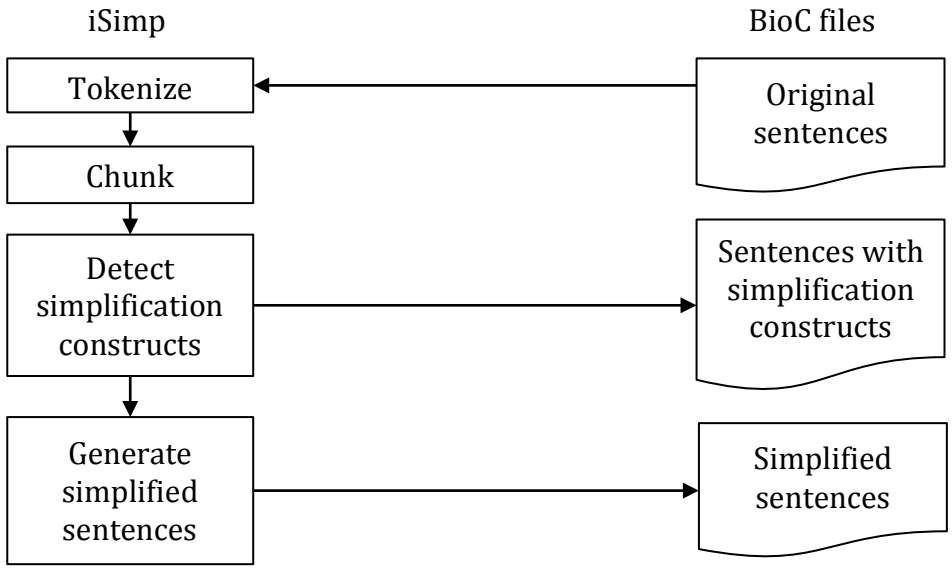

Figure 1. The workflow of iSimp

We see iSimp as a module to be used in the beginning of text mining applications. Developers can either use the webpage to submit input sentences in BioC format, or they can send POST requests to the service. The later technique will make iSimp easier to be integrated into other systems. Two types of output are provided: (1) sentences marked with simplification constructions, and (2) a list of simplified sentences, where each token is mapped back to the original text. It is often the case that new tokens will be added in simplified sentences to ensure their syntactical correctness. These new tokens will absolutely not be mapped to the original text.

BioC Format in iSimp

BioC [2] is an XML format that ensures interoperability among documents and annotations, such as part-of-speech tags, name entities, or relations. Because sentence simplification requires a

[2] http://research.bioinformatics.udel.edu/isimp/services.html

unique schema for annotation, unlike most NLP tasks, we define a BioC tag set for annotating and sharing the simplification results. We use "BioCAnnotation" to mark the simplification construct components, e.g., conjuncts and conjunctions in coordinations. We use "BioCRelation" to specify how they are related. In this way, we are able to assign roles for each component and skip over symbols like comma.

Additionally, iSimp poses a challenge to the BioC format because it also generates new simplified sentences. Such challenges were not discussed in [2]. The BioC XML file generated by iSimp contains both original and simplified sentences. Original sentences' offsets are the same as in the original text. However, simplified sentences' offsets start with the next char after the last in the original document (last document's offset + last document's length). This new collection could then be treated as the input collection for further processing in the NLP pipeline.

In order to link text in simplified sentences to that in the original sentence, we provide "equivalence" relations, which can be helpful for information extraction tasks. For example, we link "alpha-synuclein" and "phosphorylated" in both (E2) and (E3) back to (E1). Thus, only one relation <alpha-synuclein, phosphorylated> will be extracted from (E1)-(E3). This technique makes iSimp different from previous sentence simplification systems such as BioSimplify [3].

Relation extraction system

To examine the usefulness of iSimp, we designed and developed a rule-based relation extraction system. The first relation we focused on was the phosphorylation relation. We manually created a list of rules, where X is a protein or protein product. Some example rules are shown below:

1. phosphorylation of X
2. X phosphorylation
3. [$_{\text{noun phrase}}$ phosphorylated X]
4. phosphorylate (or, phosphorylates, phosphorylated, phosphorylating) X

These rules are able to match simple mentions of phosphorylation in text, however they will fail to match phosphorylation mentions in complex sentences, like the one shown below.

E4. However, the activated pAkt did not lead to [$_{\text{coordination}}$ **phosphorylation and inactivation**] of the downstream target GSK3 (PMC-2065877).

But iSimp is able to generate two simple sentences from (E4):

E5. However, the activated pAkt did not lead to **phosphorylation** of the downstream target GSK3.
E6. However, the activated pAkt did not lead to **inactivation** of the downstream target GSK3.

The first rule above can now apply on (E5) and extract <phosphorylation, GSK3>. Because the hand-crafted rules are very precise, the simplification step will only help improve the recall of the system, without hurting the precision.

We have converted the BioNLP-ST 2011 GE corpus to the BioC format for evaluation purposes. The training set, the development set, as well as the conversion script are now publicly available. The test set was not included in the release because it does not contain event annotations. Jimeno Yepes, et al. [5] discusses convention between the brat and BioC format.

Results

For others to evaluate the performance of iSimp, we provide a corpus marked with simplification constructs, using the BioC format (http://research.bioinformatics.udel.edu/isimp/corpus.html). To examine the usability of iSimp in other systems, we tested the relation extraction system on the BioNLP-ST 2011 GE task training corpus. Results show that the Precision/Recall/F-value of the phosphorylation extraction before and after simplification are 97.32/78.38/86.83 versus 97.42/81.62/88.82, respectively. Therefore, with the help of iSimp, the recall of the relation extraction system improved by 3%, while the precision stayed the same. In the ongoing work, we have observed similar improvement in the recall for other relation extraction tasks.

Conclusion

In order to participate in the BioCreative IV track 1, we adapted our previously developed system, iSimp (a sentence simplification system), to read and write BioC format files. We converted a previously annotated corpus to the BioC format to show the performance of iSimp. To emphasize the wide applicability of iSimp, we examined its impact on event extraction. We released the simplification corpus, the BioNLP corpus, and the conversion script, for others to easily judge the results and use them in comparing and designing other simplifiers.

Funding

This work was supported by the NLM of NIH [G08LM010720] and NSF [DBI-1062520].

References

1. Peng,Y., Tudor,C.O., Torii,M., Wu,C.H. and Vijay-Shanker,K. (2012) iSimp: A sentence simplification system for biomedical text. *In Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012)*, 211-216.

2. Comeau,D.C., Dogan,R.I., Ciccarese,P., Cohen,K.B., Krallinger,M., Leitner,F., Lu,Z., Peng,Y., Rinaldi,F., Torii,M., Valencia,V., Verspoor,K., Wiegers,T.C., Wu,C.H., and Wilbur,W.J. (2013) BioC: A minimalist approach to interoperability for biomedical text processing. *Database: The Journal of Biological Databases and Curation*..

3. Jonnalagadda, S. and Gonzalez, G. (2010) BioSimplify: An open source sentence simplification engine to improve recall in automatic biomedical information extraction. *AMIA Annual Symposium Proceedings*.

4. Kim, J.D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T. and Yonezawa, A. (2012) The Genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics*, 13 (Suppl 11):S1.

5. Jimeno Yepes, A., Neves, M. and Verspoor, K. (2013) Brat2BioC: conversion tool between brat and BioC. Submitted to the BioCreative IV workshop.